

# 運用機器學習於 業務員風險偵測

---

南山題目二 第三組

鄭雅綿 張培莉 潘竑叡

詹博揚 杜昕

# 保險業務員缺失行為頻傳，提前進行防範勢在必行

## 商業問題

保險業務員缺失案件層出不窮，手法日新月異，影響公司聲譽。南山人壽欲建立事前偵測機制防範缺失行為，以增加保戶的信任。

## 現行做法 與 目前痛點

隨機電訪  
→ 針對新保單、特定形式保單隨機電訪保戶  
→ 隨機抽樣可能錯失真正高風險案例

業務規則  
→ 依照過往案件及專家經驗找出可疑業務員  
→ 各種規則複雜度高，且受調查者主觀經驗影響

預測模型  
→ 建置保險業務員風險預測模型，透過風險評分聚焦高風險業務員  
→ 變數選擇過少可能導致預測能力不佳，選擇過多又可能導致過度配適，使穩定度不佳

## 專案任務

運用南山人壽現有資料，使用Python機器學習建構更精準的業務員風險模型，並兼顧模型準確性、穩定性、可解釋性

# 專案流程

## 資料前處理與 探索性資料分析

- 資料整理與整合
- 缺失值處理
- 類別變數轉換
- 探索性資料分析

## 特徵選擇與 模型建立

- 平衡樣本資料
- 特徵選擇與特徵工程
- 模型建立與交叉驗證
- 超參數調校與優化

## 模型整合與 商業解釋

- 整體資料建模
- 模型整合與評估
- 變數重要性與解釋

# 專案流程

## 資料前處理與 探索性資料分析

- 資料整理與整合
- 缺失值處理
- 類別變數轉換
- 探索性資料分析

## 特徵選擇與 模型建立

- 平衡樣本資料
- 特徵選擇與特徵工程
- 模型建立與交叉驗證
- 超參數調校與優化

## 模型整合與 商業解釋

- 整體資料建模
- 模型整合與評估
- 變數重要性與解釋

# 資料集與評估指標 (資料係經線性調整)

## 建模資料

建模資料共30,000筆，為2019年、2020年資料

- 自變數：共390個，包括以下四個面向：
  1. **基本資料與事件**：業務員所在地區、縣市、職級；工作年資、年收入、過往風險行為等
  2. **客戶面**：被保人數、保費、特殊保單數、同樣資訊保單/客戶數量、比例等
  3. **保單面**：短期保單成交、變更、失效之數量、比例；各類保單數量、金額、比例等
  4. **理賠面**：業務員理賠件數、拒絕理賠數量、金額、比例；業務員短期理賠趨勢性指標
- 應變數Y：2020年1~12月是否舞弊，為申訴資料

## 測試資料

測試資料共32,000筆，自變數為2020年資料，應變數為2021年1~3月是否舞弊

## 評估方式

### 準確性

模型預測風險最高前5%業務員，能有效捕捉到多少比例的舞弊業務員

### 穩定性

驗證資料能維持多少比例的捕捉率

### 解釋性

模型是否具備可解釋性，並具有商業洞見

# 資料前處理流程

## 讀入資料 格式調整

- 資料讀入時百分比資料、含千分位符號資料讀入時並非數字格式，經適當轉換為浮點數格式
- 縣市、職級、分群等變數含有中文資料，為分析與製圖方便皆修改為英文變數

## 資料缺失 值處理

資料缺失值包括縣市、業務員收入、Recency(最近一次發生距今月份)資料

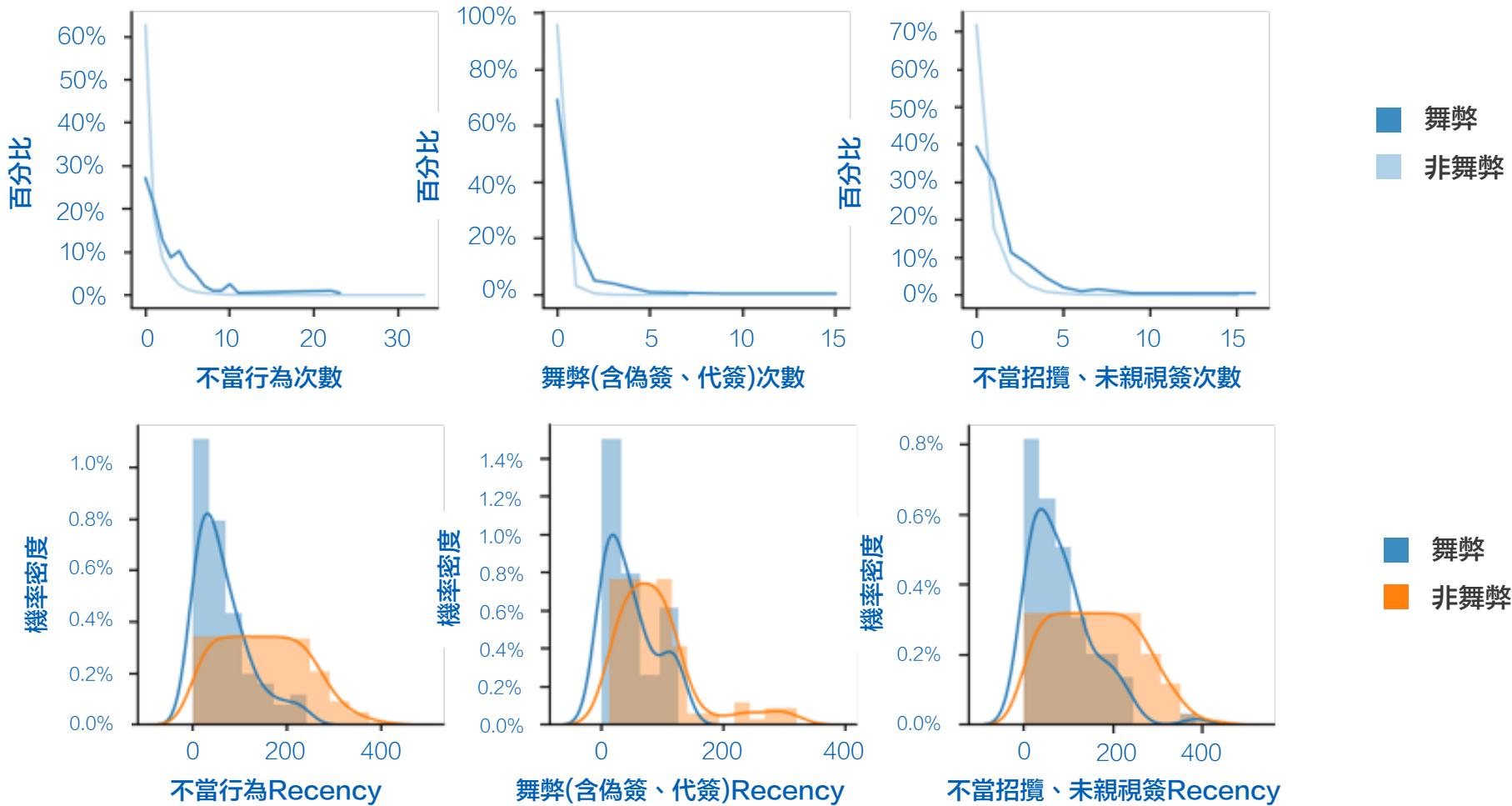
- **縣市資料**：經查詢缺失值所在分區皆位於台北，因此將縣市缺失值補上台北(TPE)
- **業務員收入**：填入該業務員所處職級、分群之中位數
- **Recency**：發生次數為 0 時沒有上一次發生紀錄，因此Recency為缺失值，補上 100,000,000 代表無上次發生紀錄

## 類別變數 轉換

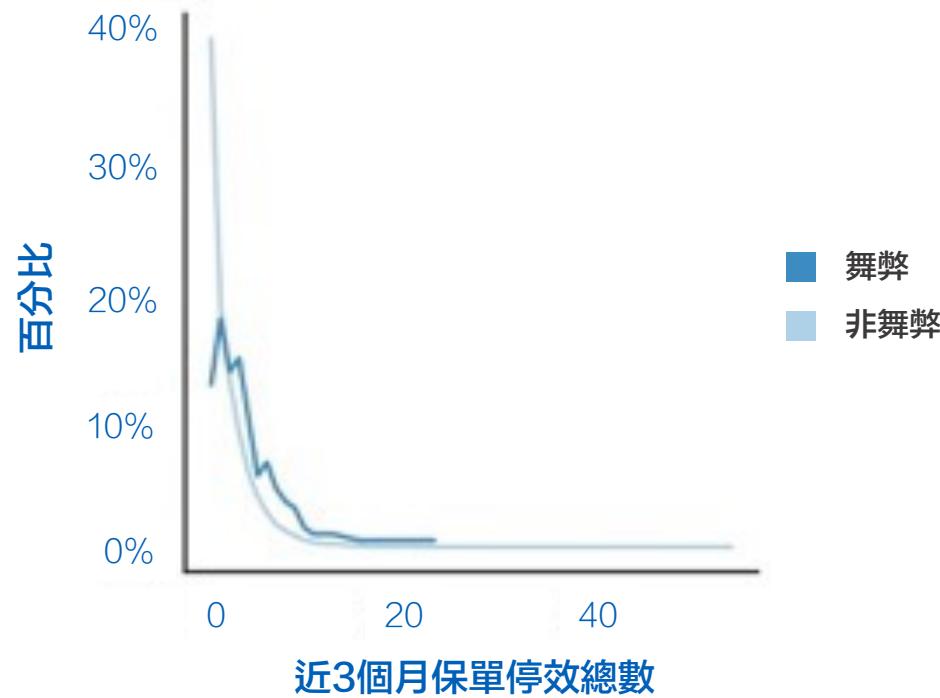
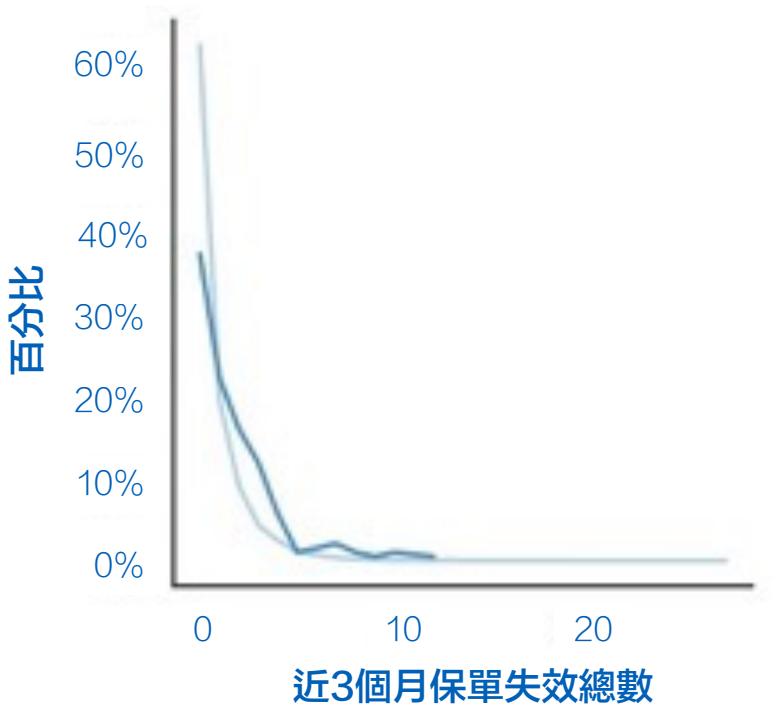
業務員職級、分群、所屬地區、所屬縣市、活動月份5個欄位為類別資料，將其轉為類別變數以利進行後續分析

資料前處理完畢後，進行探索性資料分析尋找可能潛在顯著變數

# 舞弊與否與過去舞弊次數、上次舞弊日期有顯著影響



# 舞弊與否與近期保單失效、停效總數有顯著影響



# 專案流程

## 資料前處理與 探索性資料分析

- 資料整理與整合
- 缺失值處理
- 類別變數轉換
- 探索性資料分析

## 特徵選擇與 模型建立

- 平衡樣本資料
- 特徵選擇與特徵工程
- 模型建立與交叉驗證
- 超參數調校與優化

## 模型整合與 商業解釋

- 整體資料建模
- 模型整合與評估
- 變數重要性與解釋

分割樣本  
使用交叉驗證

平衡各群樣本

特徵選擇與  
特徵工程

建立模型與  
驗證資料

超參數調校

## 分割樣本使用交叉驗證

## 平衡各群樣本

執行  
原因

妥善運用所有資料，避免過度配適。

使用  
套件

sklearn.feature\_selection  
● train\_test\_split  
● StratifiedKFold

執行  
方式

將樣本切分成數份，執行k-Fold CV，  
經反覆驗證後使用 4-Fold CV

因舞弊樣本僅佔總樣本數0.65%，使用上採  
樣將舞弊樣本增加至一定比例。

imblearn.over\_sampling  
● RandomOverSampler  
● SMOTE

超參數調整後使用RandomOverSampler  
將舞弊樣本比例提升至1 : 1

# 特徵選擇與特徵工程

執行  
原因

- 特徵選擇：可降低資料維度，縮短訓練時間，並降低模型過度配適的可能性。
- 特徵工程：增加模型變數，提升預測能力（最終因效果不顯著而未採用）

使用  
套件

- `sklearn.feature_selection.SelectFromModel`
- `sklearn.decomposition.PCA`（最終因解釋性差而未採用）
- XGBoost：`feature_importance_`

執行  
方式

1. 計算每項變數舞弊與非舞弊資料的平均、標準差，挑選Z值絕對值最大的30個變數
2. 先使用所有變數進行建模，再使用`feature_importance_`篩選最重要的40個變數
3. 使用`SelectFromModel`對每份樣本資料篩選40個變數，再挑選被重複挑選3次以上的變數

## 建立模型與驗證資料

## 超參數調校

執行  
原因

測試各種特徵選擇方式的模型準確度，  
進一步決定最終使用模型

使用  
套件

XGBoost.XGBRegressor

執行  
方式

- 使用Logistic Regression預測每筆資料的風險值，介於0~1之間
- 將資料依風險值高低排序，計算捕捉舞弊業務員準確度

將初步模型進行優化，得出預測力最佳的參數組合

sklearn.model\_selection.GridSearchCV

調整XGBRegressor的eta、max\_depth、n\_estimators、min\_child\_weight，避免過度配適，增加模型穩定性

# 專案流程

## 資料前處理與 探索性資料分析

- 資料整理與整合
- 缺失值處理
- 類別變數轉換
- 探索性資料分析

## 特徵選擇與 模型建立

- 平衡樣本資料
- 特徵選擇與特徵工程
- 模型建立與交叉驗證
- 超參數調校與優化

## 模型整合與 商業解釋

- 整體資料建模
- 模型整合與評估
- 變數重要性與解釋

# 使用全部資料重新進行建模並整合，得到最終模型

得到每個模型的最佳超參數組合後，再重新將全部樣本進行建模，得到4個最終模型，其預測風險最高前5%業務員建模資料準確度，及本組最終加成比例如下

測試資料準確度	舞弊滲透度	捕捉倍數	捕捉舞弊佔比	最終加成比例
使用所有變數	5.53%	8.51倍	42.56%	0.4
使用feature_importance選擇變數	5.40%	8.31倍	41.54%	0.3
使用Z值絕對值選擇變數	5.00%	7.70倍	38.46%	0.1
使用SelectFromModel選擇變數	5.27%	8.10倍	40.51%	0.2
加成後模型精確度	6.07%	9.33倍	46.67%	

使用加成後模型預測建模資料，風險最高前5%捕捉到46.67%的舞弊業務員，優於業主使用傳統計分卡的預測結果31.8%。

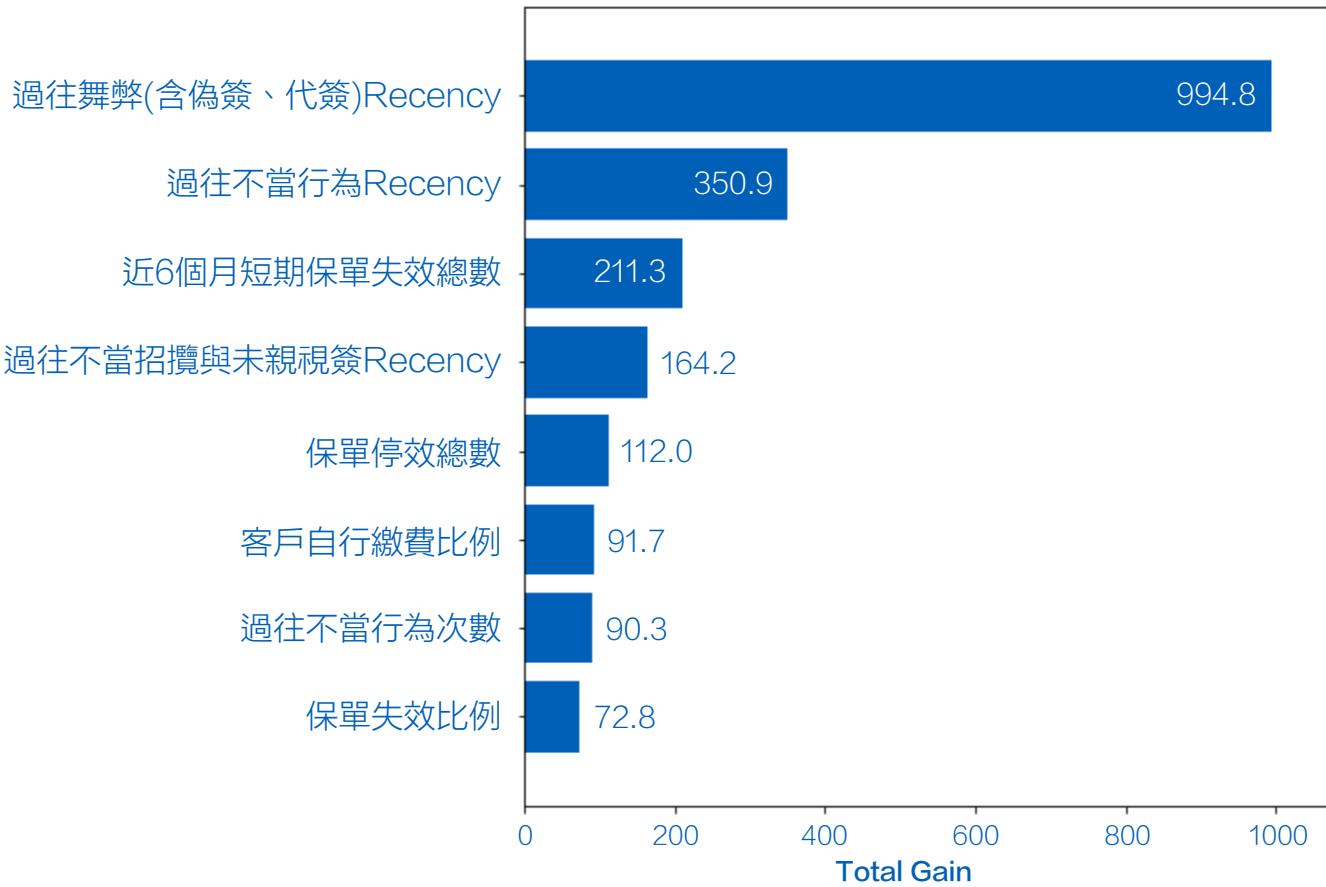
# 測試資料結果：風險最高的前5%業務員捕捉到36.36%舞弊

在預測資料中，風險最高前5%業務員捕捉到36.36%舞弊業務員，優於業主使用傳統計分卡的預測結果29.1%。

百分比	累積資料數	累積舞弊資料數	舞弊滲透度	捕捉倍數	捕捉舞弊佔比	捕捉舞弊累積佔比
5%	1,600	20	1.25%	7.27倍	36.36%	36.36%
10%	3,200	27	0.44%	2.55倍	12.73%	49.09%
20%	6,400	32	0.22%	1.27倍	12.73%	58.18%
30%	9,600	39	0.22%	1.27倍	12.73%	70.91%
40%	12,800	41	0.06%	0.37倍	3.64%	74.55%
50%	16,000	47	0.16%	0.91倍	9.09%	83.64%
60%	19,200	51	0.13%	0.73倍	7.27%	92.73%
70%	22,400	52	0.03%	0.36倍	3.64%	94.55%
80%	25,600	54	0.03%	0.18倍	1.82%	98.18%
90%	28,800	54	0.00%	0.00倍	0.00%	98.18%
100%	32,000	55	0.03%	0.18倍	0.00%	100%

# 過往舞弊資料、保單停效失效、客戶自行繳費為最顯著變數

根據XGBoost前8項重要變數 (Total Gain)



如同探索性資料分析的結果，舞弊資料與非舞弊資料的過往舞弊經驗有極大差異，模型建立時也將過往舞弊經驗視為最重要變數。

業務員可能透過偽造簽名等方式，私自將要保人保單送至停效階段，或替要保人申請未發生的保險申請等方式進行舞弊。

#### 過往舞弊(含偽簽、代簽)Recency

#### 過往不當行為Recency

近6個月短期保單失效總數

#### 過往不當招攬與未親視簽Recency

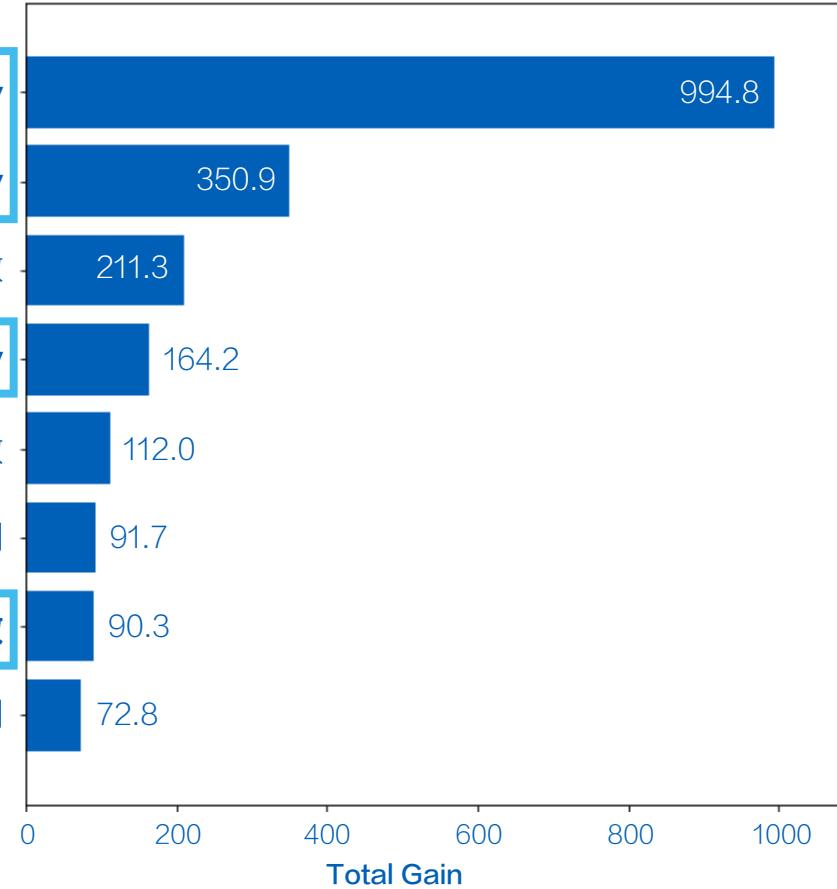
保單停效總數

客戶自行繳費比例

#### 過往不當行為次數

保單失效比例

根據XGBoost前8項重要變數 (Total Gain)



保單若在應繳費期間無繳費，保單會進入停效階段，待一定期間後才會轉為失效保單。

業務員可能利用保單停效規定，將客戶保單價值移轉，從事自身投資或是償債行為等，待價值回收再移轉回客戶保單。因此若保單有停效、失效等情形，可能是業務員從事舞弊的徵兆。

過往舞弊(含偽簽、代簽)Recency

過往不當行為Recency

近6個月短期保單失效總數

過往不當招攬與未親視簽Recency

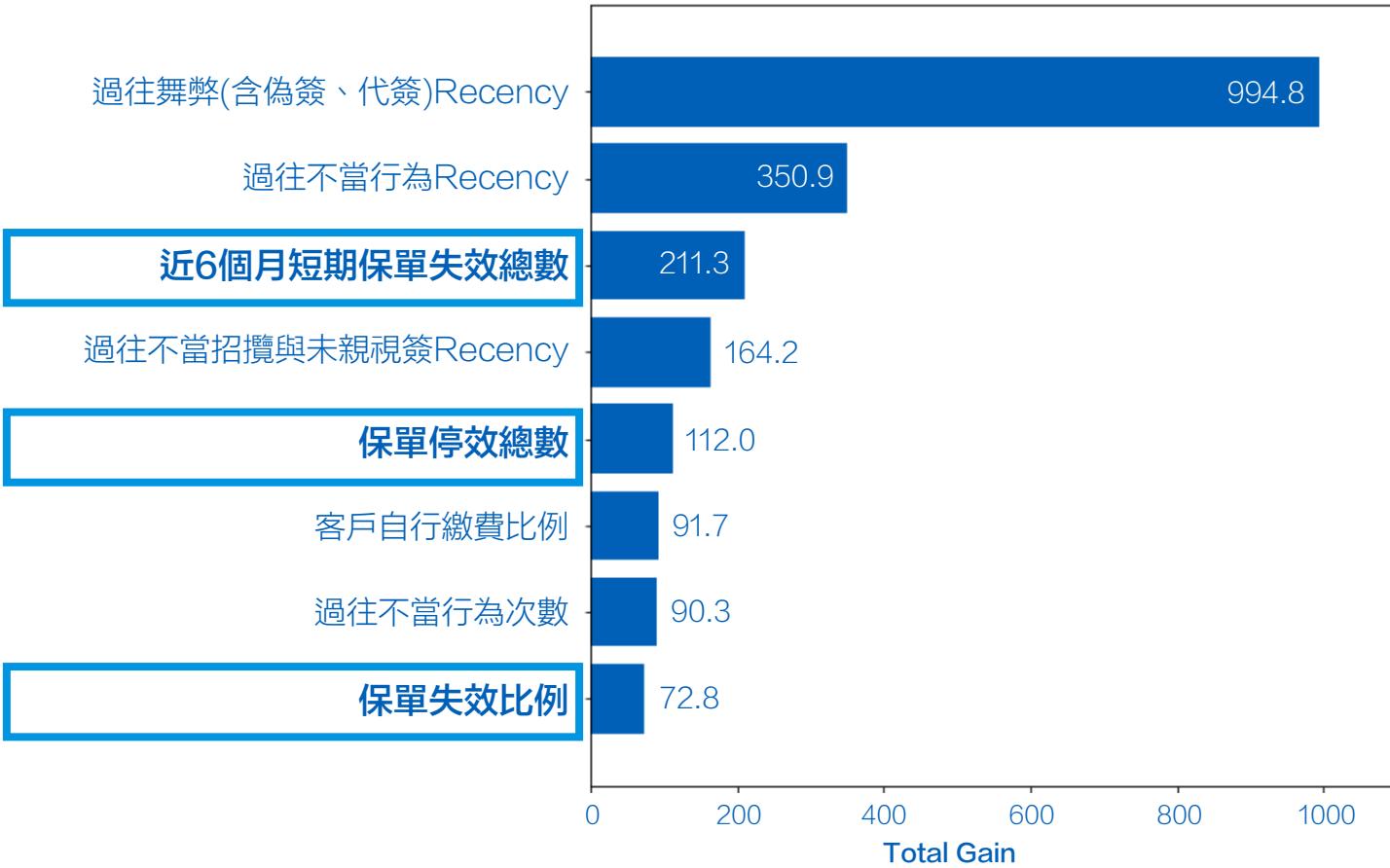
保單停效總數

客戶自行繳費比例

過往不當行為次數

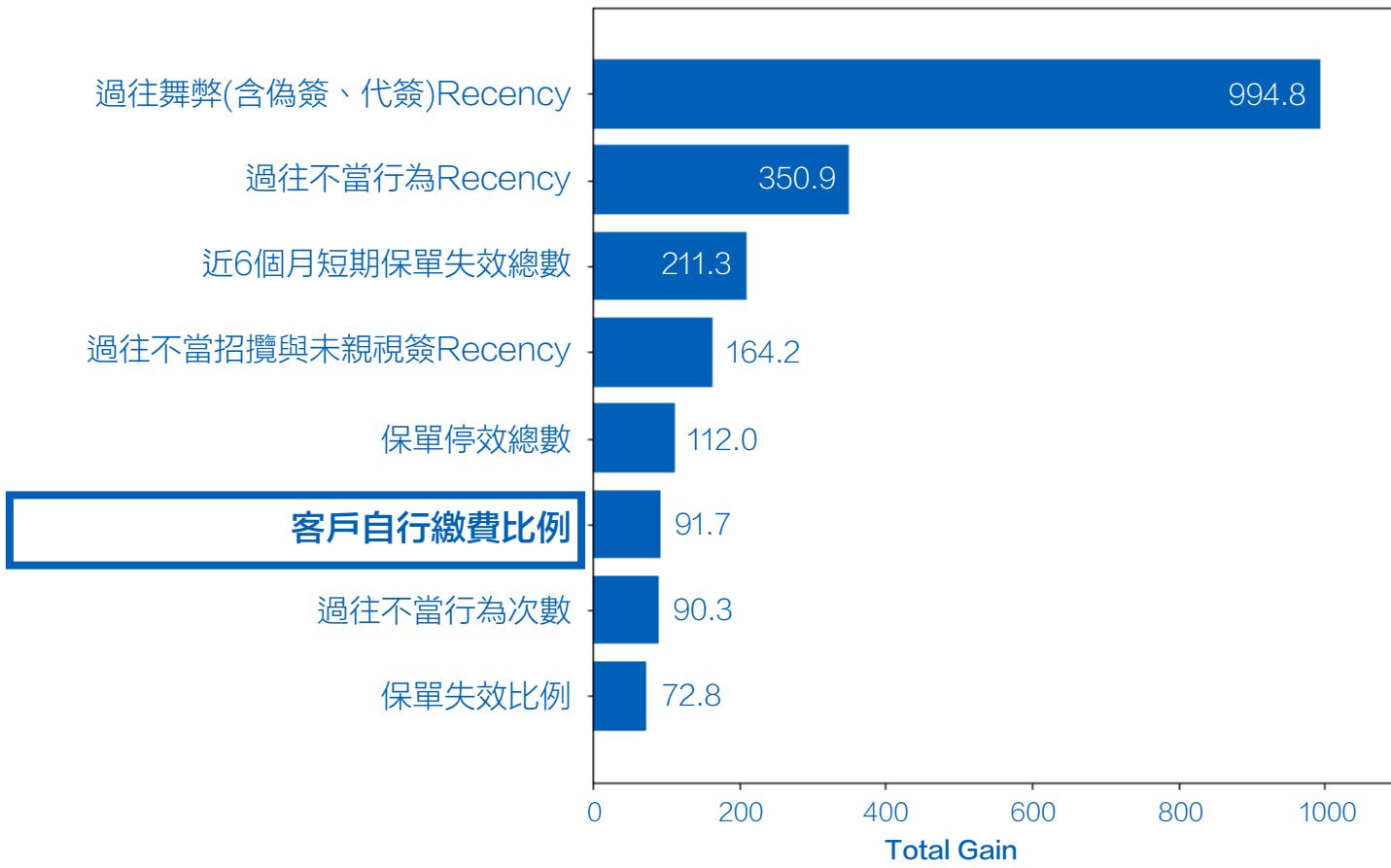
保單失效比例

根據XGBoost前8項重要變數 (Total Gain)



根據業主說明，客戶自行繳費可能聽從業務員指示自行將款項匯入，待一段期間後再向客戶說明更換帳戶，進而將款項匯入自己的帳戶中，從事舞弊行為。根據模型顯示，客戶自行繳費比例亦為模型重要變數之一。

根據XGBoost前8項重要變數 (Total Gain)



# 感謝聆聽

---

南山題目二 第三組

鄭雅綿 張培莉 潘竑叡

詹博揚 杜昕

# 附錄：組員分工

雅綿	聯絡業師教授、安排會議時間、PCA、特徵變數解釋
培莉	會議記錄、PCA、新增變數、特徵變數解釋
竑叡	PCA、特徵變數解釋
博揚	探索性資料分析、超參數調校
杜昕	資料前處理探索性資料分析、特徵選擇、模型建立與整合、簡報製作與影片錄製